I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

CHRISTOPHER SOGHOIAN*

# The Problem of Anonymous Vanity Searches

**Abstract:** This paper explores privacy problems stemming from search behavior conducted using public search engines. Specifically, it exposes problems related to unintentional information leakage through a vanity search, which is a search for information about oneself. This article begins by discussing recent events that have made this problem extremely salient. It introduces a number of existing technologies, such as Tor and TrackMeNot, which aim to protect users' privacy online and explains how each of these programs fails to protect users against the specific risks related to self-search. This article highlights the inherent information asymmetry in the relationship between search engines and their users that makes it almost impossible to create cover traffic good enough for privacy-desiring users to blend their own searches into. The article concludes by exploring other avenues for protecting user privacy online.

* Chris Soghoian is at The School of Informatics, Indiana University, Bloomington. The author can be reached at csoghoia@indiana.edu. The author was partially funded by a scholarship from Google and the Hispanic College Fund. This article is not part of any research conducted as part of an internship the author conducted at Google. It is not sponsored or authorized by Google and does not reflect or take into account the company's views. This article was written using publicly available information and is based upon the personal views of the author. The author would like to extend many thanks to Kelly Caine, L. Jean Camp, Allan Friedman, Markus Jakobsson, Sid Stamm and Katherine Townsend for their helpful comments.

## I. INTRODUCTION

On August 4, 2006, America Online, Inc. (AOL) publicly released the search records of 650,000 of its users. The stated goal of this release was to aid the research community at large.[1] In an effort to protect user privacy, the records were "pseudonymized" by replacing each individual customer's account I.D. and computer network address with unique random numbers. While this severed the connection between the search records and AOL account information, such as a user's name and address, the ability to link an individual's searches over multiple sessions remained the same. Almost a quarter of all Internet users engage in the common practice of searching for their own name or online nickname (which can include an email address, instant messenger ID, or MySpace address) on the Internet.[2] Due to this behavior, often called a vanity search or self-Googling,[3] it was possible for journalists from The New York Times to reveal the identity of user 4417749 to be Thelma Arnold, a 62-year-old widow from Lilburn, Georgia. This discovery was made by linking together all of her vanity searches contained in AOL's pseudonymized records.[4]

Court papers filed on January 18, 2006 by the U.S. Department of Justice revealed that Google refused to cooperate with a previous year's subpoena for user search records.[5] Data requested included one million randomly indexed webpage addresses and records for all searches performed during a one-week period. In its refusal, Google cited the privacy rights of its customers and the risk of revealing

---

[1] Dawn Kawamoto, "AOL Apologizes for Release of User Search Data," *CNETNews.com*, August 7, 2006, http://www.news.com/2100-1030_3-6102793.html (accessed November 8, 2007).

[2] Deborah Fallows, "Search Engine Users," *Pew Internet & American Life Project* (2005): 1-2, http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf (accessed November 8, 2007).

[3] Daniel Dasey, "A Quick Self-Google Once a Day to Guard Your Reputation," *Sydney Morning Herald*, May 23, 2004, http://www.smh.com.au/articles/2004/05/22/1085176043551.html (accessed November 8, 2007).

[4] Michael Barbaro and Tom Zeller Jr., "A Face is Exposed for AOL Searcher No. 4417749," *New York Times*, August 9, 2006, http://www.nytimes.com/2006/08/09/technology/09aol.html (accessed November 8, 2007).

[5] See *Gonzales v. Google, Inc.*, 234 F.R.D. 674 (N.D. Cal. 2006).

company trade secrets.[6] Lawyers from the Department of Justice argued that they needed the information to prepare their defense of the 1998 Child Online Protection Act, a law that the courts had previously struck down, stating that it was too broad and had the potential to prevent adults from accessing legal pornography sites. Both Yahoo and MSN silently complied with the request, and if Google had not publicly refused to do so, it is likely that the subpoenaed information would have been handed over without public notification.

When testifying in front of a U.S. Senate panel on September 19, 2006, U.S. Attorney General Alberto Gonzales urged the Senate to pass legislation requiring Internet Service Providers (ISPs) to keep two years' worth of detailed log data on their customers' online browsing habits.[7] He stated that the growing threat of child pornography made it necessary to keep such information for subsequent law enforcement investigations.[8]

Search engine logs are increasingly being sought in other cases. In one case, a murder suspect's search records were produced in court to prove that he had searched for the words "neck," "snap," and "break" before killing his wife.[9] Others have speculated that it is only a matter of time before search records are subpoenaed in civil cases, including divorces.[10]

---

[6] Declan McCullagh, "Google, Feds Face Off Over Search Records," *CNETNews.com*, March 14, 2006, http://news.com.com/Google,+feds+face+off+over+search+records/2100-1028_3-6049262.html (accessed November 8, 2007).

[7] Alberto Gonzales, "Combating Child Pornography by Eliminating Pornographers' Access to the Financial Payment System," (Statement before Senate Committee on Banking, Housing and Urban Affairs, September 19, 2006) http://banking.senate.gov/_files/gonzales.pdf (accessed November 8, 2007).

[8] "Gonzales Wants New Web Rules," *CBS News*, September 19, 2006, http://www.cbsnews.com/stories/2006/09/19/politics/main2023209.shtml (accessed November 8, 2007).

[9] Julia Lewis, "Petrick Googled 'Neck,' 'Snap,' Among Other Words, Prosecutor Says," *WRAL News*, November 13, 2005, http://www.wral.com/news/local/story/121729/ (accessed November 8, 2007).

[10] Declan McCullagh, "FAQ: When Google is Not Your Friend," *CNETNews.com*, February 3, 2006, http://news.com.com/FAQ+When+Google+is+not+your+friend/2100-1025_3-6034666.html (accessed November 8, 2007).

## II. PROTECTING PRIVACY ONLINE

Shortly after the AOL search records were released, the Electronic Frontier Foundation (EFF) released a set of recommended best practices for safe search behavior on the Internet.[11] These include:

- Do not put personally identifying information in your search terms;
- Do not use your ISP's search engine;
- Do not login to your search engine or related tools;
- Block "cookies" from your search engine;
- Vary your Internet Protocol (IP) address;
- Use web proxies and anonymizing software, such as Tor.[12]

While its first piece of advice is never to search for information on oneself, the EFF seems to accept that even though conducting vanity searches may be risky, people will do them anyway.[13] The EFF suggests that such users should heed the rest of its advice, or at the least, run those sensitive searches from another computer than the one used for normal search activity.[14]

This article argues that certain terms placed together in a search log, even without identifying information such as an IP address, cookie, or user account, still reveal far too much information about the user. There is a considerable difference between surfing the Internet privately (such as by using Tor to keep the network address of your computer hidden from the websites you visit) and not revealing your identity through your search behavior. Had the AOL customers who were identified turned off their cookies and used an anonymizing proxy such as Tor, it is extremely unlikely that The New York Times would have been able to link together their individual, non-vanity

---

[11] Peter Eckersley and others, "Six Tips to Protect Your Online Search Privacy," *Electronic Frontier Foundation*, 2006, http://www.eff.org/wp/six-tips-protect-your-search-privacy (accessed November 8, 2007).

[12] Roger Dingledine, Nick Mathewson, and Paul Syverson, "Tor: The Second-Generation Onion Router," in *Proceedings of the 13th USENIX Security Symposium* (San Diego, CA: USENIX, 2004) http://www.usenix.org/events/sec04/tech/full_papers/dingledine/dingledine.pdf (accessed November 8, 2007).

[13] Eckersley and others, "Six Tips to Protect Your Online Search Privacy."

[14] Ibid.

searches from the released log data. The analysis of the search logs would reveal the fact that someone was searching for "Britney Spears," but the identity of that someone would remain secret.

Some search terms are so sensitive that their mere presence in logs is likely to cause alarm to the user. Examples of this type of query may include the combination of a user's name or online nickname with terms such as "HIV," "rape," "sexual harassment," "sex offender," and "homosexual." After-the-fact analysis of search logs will not confirm that the subject of the sensitive search query was also the person issuing the search. Nevertheless, the mere fact that someone was looking for such combinations is interesting and extremely sensitive in and of itself.

## III. A CONFLICT OF INTEREST IN THE ADVERTISING BUSINESS

The major search engines depend upon advertising for their revenue. Google provides free email, free wireless Internet access,[15] and mobile phone-based GPS mapping services,[16] all so that it can offer targeted advertisements to customers most likely to respond to the ads. If Google can build a higher-quality data set of customer information, it can charge more per advertisement, while also gaining a significant market advantage over other search engines.

Some have claimed that loss of privacy on the Internet has allowed merchants to perform more effective price discrimination because it allows them to determine exactly how much each customer is willing to pay for a product.[17] While Google's search engine and email products are clear market leaders in terms of quality and functionality, users are not given the choice between a subscription charging (yet privacy preserving) model and the more common advertising supported (and privacy denying) model. Instead users are given the binary choice of either using the products, with the advertisements and potential intrusions into their privacy, or not using them at all. If users wish to attempt to preserve their privacy and avoid advertising, they must take matters into their own hands. They cannot ask the search

---

[15] Google WiFi, "Google WiFi Mountain View," https://wifi.google.com/support (accessed November 8, 2007).

[16] Alex Medina, "Get Lost!," *Official Google Blog*, November 9, 2006, http://googleblog.blogspot.com/2006/11/get-lost.html (accessed November 8, 2007).

[17] Andrew Odlyzko, "Privacy, Economics, and Price Discrimination on the Internet," July 27, 2003, http://www.dtc.umn.edu/publications/reports/2003_13.pdf (accessed November 8, 2007).

engines to take care of this for them, nor can they financially compensate the search engines for the potential lost revenue due to the loss of accurate user data.

Users have struck a Faustian bargain of sorts with the major search engines. They seem to be willing to put up with advertising and a wholesale loss of privacy, assuming that they are even aware that it is happening, for free access to the services that search engines offer. Just as a generation of American college students have shared their personal financial data with credit card companies in order to get a free t-shirt or pizza,[18] Internet users seem to be engaged in a similar mass exchange of their personal information for access to accurate search results.

Most Internet users place blind trust in search engines to present only the best or most accurate, unbiased results.[19] In the vast majority of cases, users are not seeking out advertisements when they go to a search engine. They are instead trying to locate the results that most accurately match their search query. Three out of five searchers do not realize that search engines are compensated for some of the results in their listings,[20] but only one in six Internet users are able to tell the difference between paid advertisements and unbiased search results. Thus, the line between paid advertisements and genuine search results can be extremely blurry, and some search engines take advantage of this feature to increase the click-through rate of their advertisements.[21]

The search engines must be very careful to ensure that advertisements do not become too intrusive or disruptive. In particular, flash-based or JavaScript advertisements that take control of a user's screen can be irritating enough that some users seek ways to disable them.[22] Pop-up advertisements have proven to be annoying

---

[18] Lynn O'Shaughnessy, "Credit Cards Offer College Students Early Danger Lesson," *San Diego Union-Tribune*, September 24, 2006, http://www.signonsandiego.com/uniontrib/20060924/news_lz1b24oshaugh.html (accessed November 8, 2007).

[19] Leslie Marable, "False Oracles: Consumer Reaction to Learning the Truth About How Search Engines Work: Results of an Ethnographic Study," *Consumer WebWatch*, June 30, 2003, http://www.consumerwebwatch.org/pdfs/false-oracles.pdf (accessed November 8, 2007).

[20] Ibid.

[21] Associated Press, "Users Confuse Search Results, Ads," *Wired News*, January 23, 2005, http://www.wired.com/news/culture/0,1284,66374,00.html (accessed November 8, 2007).

[22] Tom Spring, "Net Watchdog: The Most Annoying Online Ads," *PC World*, September 26, 2006, http://pcworld.com/article/id,127207-page,1-c,topics/article.html (accessed November 8, 2007).

enough that over 80% of users with high-speed Internet connections now employ technology to block them – a 100% increase during the past two years.[23] Even Google's relatively unobtrusive text-based advertising has inspired a number of avoidance technologies.[24] Savvy people have been using technologies that allow them to skip advertisements for a number of years. Use of such programs by a tiny percentage of users does not have a significant or even measurable impact upon the search engines' revenue. However, consider the cases of television advertisement skipping with ReplayTV/Tivo and the P2P file-sharing technologies made mainstream by Napster. When this kind of easy-to-use technology enables an average user to bypass the copyright/advertising systems in place, it threatens to destabilize the entire business model upon which companies' revenue streams are built.

Google would no doubt prefer that each user sign in to one of the company's services before searching. In such cases, Google is able to collect far more data to link searches, advertising clicks, and browsing habits with an individual person across multiple sessions and from different computers. Likewise, simply letting Google store a persistent cookie on one's computer allows the company to achieve its goals, albeit with a significantly lower quality of user data. It is not surprising that each of the EFF's private searching recommendations threaten Google's bottom line, should these measures be employed by the masses. As of September 2006, the highest-valued search terms in Google's AdWords advertising system are related to medical class action lawsuits and other legal problems.[25] Yet, these are the same kinds of searches that users will likely want to protect from prying eyes, or worse, a subpoena after the fact. One of the main goals of this article is to highlight this relationship: that Google's ability to serve fine-grained advertising (and thus achieve higher revenues) directly

---

[23] Thomas Claburn, "Consumer Use of Ad-Blocking Technology Doubles," *Information Week*, December 5, 2006, http://www.informationweek.com/internet/showArticle.jhtml?articleID=196601694 (accessed November 8, 2007).

[24] CustomizeGoogle: Improve Your Google Experience, http://www.customizegoogle.com/ (accessed November 8, 2007); Userscripts.org, "Hide Google Adsense Ads," http://userscripts.org/scripts/show/675 (accessed November 8, 2007).

[25] John Battelle, "Highest Paying AdWords," *John Battelle's Searchblog*, March 26, 2006, http://battellemedia.com/archives/002444.php (accessed November 8, 2007); CyberWyre, "Updated: Highest Paying AdSense Keywords," March 23, 2006, http://www.cwire.org/2006/03/23/updated-highest-paying-adsense-keywords/ (accessed November 8, 2007).

competes with the methods by which a user can attempt to achieve anonymity and preserve his or her online privacy.

## IV. TRACKMENOT

Shortly after the AOL incident became public, New York University researchers Daniel C. Howe and Helen Nissenbaum released a tool named TrackMeNot (TMN)[26] that aims to protect user search privacy. Their tool is an extension to the Firefox web browser and initiates randomized search queries in the background to a number of commercial search engines. These searches, issued over random periods of time, aim to lose the user's real searches in a cloud of "ghost queries" and as the authors describe, "significantly [increase] the difficulty of aggregating such data into accurate or identifying user profiles."[27]

### A. THE DIFFICULTY OF FAKING TRAFFIC

The current version of TMN begins only with a small seed list taken from lists of most frequent search terms published by the search companies. Using these terms as seeds, each TMN client dynamically evolves its query list by parsing likely search terms from the results of each query and swapping these back into its "Current-Query" list.[28]

There is very little risk to the user if someone else learns that they are searching for the hot topic, or immensely popular search term of the day. Any hot topic is sought by millions of other users, thus one is unlikely to be embarrassed or suffer otherwise if searches for those terms become public. As a concrete example, one wants privacy when searching for information on breast cancer, but not when searching on Britney Spears. TMN creates cover traffic, or "ghost queries" as the authors describe them, by submitting queries to search engines containing terms from its Current-Query list at random intervals.

The actual level of plausible deniability provided to users could prove to be rather problematic due to the fact that users do not submit their own legitimate queries at random intervals. They tend to send

---

[26] Daniel C. Howe and Helen Nissenbaum, "TrackMeNot," http://mrl.nyu.edu/~dhowe/TrackMeNot (accessed November 8, 2007).

[27] Ibid.

[28] Daniel C. Howe and Helen Nissenbaum, "TrackMeNot FAQ," http://mrl.nyu.edu/~dhowe/TrackMeNot/faq.html (accessed November 8, 2007).

multiple searches over a short period of time, followed by longer periods of web-browsing. It should be fairly easy for the search engine companies to focus on users' real search activity, by filtering out all searches that match the behavior exhibited by TMN.

The major search engines have extremely accurate data on the search frequency of various terms. Portions of this information, albeit generalized to remove specific traffic figures, are even made public through services like Google Trends[29] and Google Zeitgeist.[30] TMN will find itself sticking out if the search queries it issues significantly diverge from the norm and deviate from the standard frequency of search terms in the population at large, or even those other users in one's geographic location. Users often do not find the information they want from the first clicked-upon link that they reach via the search engine. A rather limited eye-tracking study by the firm Enquiro found that users returned to the search page and then clicked on additional items listed by the search engine over 49% of the time.[31] This behavior, dubbed "pogo sticking," suggests that the first link returned is often not enough to meet the user's needs. Researchers analyzing the AOL search data have also noted a strong tendency for people to refine their searches. If the first page of results does not deliver what they are after, they will refine the search terms in an effort to produce better results.[32]

Google, Yahoo, and other search engines have large amounts of accurate data for this kind of behavior and guard it carefully as a trade secret. Due to the fact that the TMN developers do not have access to detailed user click data, it will be exceedingly difficult for them to attempt to accurately mimic genuine user habits with automated searches. As a result, the TMN initiated searches will diverge from the norm of real user behavior, making it easy for the search engines to identify and filter out TMN's searches.

---

[29] Google Trends, "About Google Trends," http://www.google.com/intl/en/trends/about.html (accessed November 8, 2007).

[30] Google Press Center, "Zeitgeist: Search Patterns, Trends, and Surprises," http://www.google.com/press/zeitgeist.html (accessed November 8, 2007).

[31] Gord Hotchkiss, "Tales of Pogo Sticks, Bouncy SERPS and Sticky Pages," *Search Engine Guide*, September 11, 2006, http://www.searchengineguide.com/hotchkiss/2006/0911_gh1.html (accessed November 8, 2007).

[32] Geoffrey Faivre-Malloy, "AOL Search Data," *Search Engine Optimizer Blog*, August 22, 2006, http://www.seomoz.org/blog/aol-search-data (accessed November 8, 2007).

## B. UNINTENDED CONSEQUENCES

Search engines such as Google monitor the click-through-rates (CTR) of the advertisements they display in web pages containing search results and elsewhere. Advertisements that consistently suffer from low CTR will be punished, and in time, no longer be displayed to the user, no matter how much the advertiser pays per click.[33] In attempting to mask the user's real search behavior, TMN, in fact, will be inadvertently performing a specific kind of attack against Google's advertisers: impression spam.[34]

In addition to performing this attack, TMN will no doubt stand out due to the fact that its ghost queries will have a 0% CTR. If the TMN developers attempted to fix this behavior by modifying their program to click on advertisements at random, approximating the rate of real users, the developers would soon find themselves engaged in a different, yet equally unfriendly, behavior with respect to Google: mass click-fraud. Advertisers are charged each time one of their advertisements is clicked. Thus, each phantom search and click performed by TMN would result in a financial loss for web advertisers.

Google's terms of service clearly state that the kind of behavior that TMN engages in is forbidden.[35] Users may not send automated queries of any sort to Google's system without express permission in advance from Google. At the very least, Google would be perfectly within its rights to terminate the accounts of customers who install TMN.[36] Given that one can use Google's search engine without an account, this account termination is probably not a problem for the vast majority of Internet users. However, for those eight-million-plus Gmail users who have entrusted Google with their email data, being kicked off the service could prove to be extremely detrimental.[37]

---

[33] Google AdWords, "Learning Center," http://www.google.com/adwords/learningcenter/text/18754.html (accessed November 8, 2007).

[34] Rob McGann, "Impression Spam Worries Google Advertisers," *ClickZ News*, February 24, 2005, http://www.clickz.com/showPage.html?page=3485386 (accessed November 8, 2007).

[35] Google Privacy Center, "Google Terms of Service for Your Personal Use," http://www.google.com/terms_of_service.html (accessed November 8, 2007).

[36] Ibid.

[37] Saul Hansell, "In the Race with Google, It's Consistency vs. the 'Wow' Factor," *New York Times*, July 24, 2006, http://www.nytimes.com/2006/07/24/technology/24yahoo.html (accessed November 8, 2007).

The legality of TMN's techniques is not clear, and use of it and similar tools may expose users to legal risks.[38] The tort claim of "trespass to chattel" has been successfully used by service providers against unwanted, automated, resource-hogging Internet tools in the past.[39] While the laws surrounding click-fraud are not yet mature, Google has brought a number of cases to trial against people for attempting to defraud their AdSense advertising system.[40] The legal issues that surround TMN's behavior are beyond the scope of this paper, although they merit thorough analysis. Researching them is left as an exercise to the legally-inclined reader.

### C. THE LACK OF A FEEDBACK LOOP

One of the major problems for TMN's developers may be that it is difficult for them to measure their success. The major search engines have very little incentive to share information with TMN. If the search engines are indeed able to detect the presence of TMN, or worse, filter out the automated searches from legitimate queries, TMN's authors and users will never know. Google and others will be able to successfully use this data, and potentially gain even more value from it, with the knowledge that it is data that users value enough to go out of their way to protect. Needless to say, if the National Security Agency or some other government agency were to obtain Google's log data, they could also perform the same analysis, silently, of course.

---

[38] The additional overhead of a single copy of TMN running on a user's computer would not cause a noticeable increase in the resources required for Google to respond to the ghost queries. However, the collective effect of 20,000 copies of TMN running on thousands of different computers could potentially impair the regular functioning of Google's servers. The legal implications of this, and in particular, who Google could go after (individual TMN users, or the developers of the system) are unclear.

[39] Pamela Samuelson, "Unsolicited Communications as Trespass?" *Communications of the ACM* 46, no. 10 (2003): 15-20, http://www.ischool.berkeley.edu/~pam/papers/acm_vol46_p15.pdf (accessed November 8, 2007); Dan L. Burk, "The Trouble with Trespass," J. SMALL & EMERGING BUS. L., 3 (2000): 27, http://www.isc.umn.edu/research/papers/trespass-ed2.pdf (accessed November 8, 2007).

[40] Ben Elgin, "The Vanishing Click-Fraud Case," *BusinessWeek*, December 4, 2006, http://businessweek.com/technology/content/dec2006/tc20061204_923336.htm?chan=technology_technology+index+page_more+of+today%2-7s+top+stories (accessed November 8, 2007); Davis A. Vise, "Clicking to Steal," *Washington Post*, April 17, 2005, http://www.washingtonpost.com/wp-dyn/articles/A58268-2005Apr16.html (accessed November 8, 2007).

The absence of a feedback loop will make it extremely difficult for the TMN creators to evolve their technology, as they will be denied knowledge of which particular behavior allows their program to be identified. If TMN is widely adopted and actually becomes a significant burden on the search engines' network resources, the search engines may have to adopt a more active approach by banning users of the product. In the past, Google has blocked technologies such as Tor when known Tor IP addresses were submitting too many queries.[41] In this case, Google offered two options to users wishing to search from a Tor IP address: solve a CAPTCHA test,[42] which requires users to type the letters of a distorted image into a form, thus confirming that the user is real and not a computer program; or be blocked. One can easily imagine Google deploying a similar system for all the queries sent by a TMN user, both automated and legitimate, if they are able to easily detect the presence of the extension through search log analysis. This could instantly break TMN, unless of course, users are willing to solve a CAPTCHA for every fake search submitted by the extension, a rather unlikely scenario.

### D. INFORMATION ASYMMETRY

The relationship between search engines and their customers can be seen as a classic case of information asymmetry.[43] The search engines know how often each word is searched for, how many searches per session any one user issues on average, how much time there is between sessions and individual searches, how many advertisements are clicked on per session, and how often advertisements are expected to be clicked by a given user. Furthermore, they keep the vast majority of this information to themselves because their competitors, those actors trying to actively defraud them (e.g. those committing click-fraud), and those users trying to hide their search behavior, would all love to have this data.

---

[41] Danny Sullivan, "More on Google & Blocking Privacy Proxies," *Search Engine Watch*, September 8, 2006, http://blog.searchenginewatch.com/blog/060908-080437 (accessed November 8, 2007).

[42] Luis von Ahn and others, "CAPTCHA: Using Hard AI Problems For Security," in *Advances in Cryptology – EUROCRPYT 2003: International Conference on the Theory and Applications of Cryptographic Techniques* (Heidelberg, Germany: Springer-Berlin, 2003) http://www.cs.cmu.edu/~biglou/captcha_crypt.pdf (accessed November 8, 2007).

[43] George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics* 84 no. 3 (1970): 488–500, http://ideas.repec.org/a/tpr/qjecon/v84y1970i3p488-500.html (accessed November 8, 2007).

It is because of this huge gap in information that technologies such as TrackMeNot are doomed to failure. They lack accurate Internet behavior data that is essential in helping a user mask her searches in a cloud of effective cover traffic. Their attempts to maintain privacy without the necessary information on what cover traffic *should* look like may cause their users to stand out even more than they would if they had not attempted to evade the watchful eyes of search engines in the first place.

## V. TOR

> Tor is a network of virtual tunnels that allows people and groups to improve their privacy and security on the Internet . . . . Individuals use Tor to keep websites from tracking them and their family members, or to connect to news sites, instant messaging services, or the like when these are blocked by their local Internet providers.[44]

Tor allows users to mask the link between their own network activity (search behavior, web browsing, or instant messaging) and any logs kept by webmasters, or worse, intrusive governments. Servers receiving web requests see only a Tor exit node and are unable to learn the actual IP addresses of the users initiating queries.[45]

Use of the Tor anonymity proxy to sever the link between a user and an identifying IP address (and in tandem, disabling cookies in the browser) restricts the search engines' abilities to link an individual's searches together. Users who are solely concerned with protecting their search behavior against log analysis by the search engines do not necessarily have to use Tor – any proxy server will work.

The list of Tor exit nodes is public, and as Google's past behavior of selectively blocking Tor servers at times has demonstrated, Google subscribes to this list. Presumably, if a user issues a non-common search via a Tor server, even with cookies turned off, and then a few minutes later issues a refined but similar search query via a different Tor exit node, Google can link those two searches together. While the link between the two queries is not certain, as is the case when cookies are present, it still has the potential to reveal information that the user

---

[44] The Tor Project, Inc., "Tor: Overview," October 24, 2007, http://www.torproject.org/overview.html.en (accessed November 8, 2007).

[45] Dingledine, Mathewson, and Syverson, "Tor: The Second-Generation Onion Router."

expected to remain private. This technique only works for uncommon search queries. Yet, as explained earlier in this article, these are often the searches that users wish to protect the most.

### A. WHY TOR ALONE CANNOT PROTECT VANITY SEARCHES

The combination of an anonymizing proxy, such as Tor, and a browser session without cookies does much to protect user search activity on the Internet, and in particular, the linking of multiple queries during one or more sessions. As was explained earlier in this article, a search for one's own name combined with culturally or politically delicate terms can be extremely sensitive search information. While Tor will deny the search engine the knowledge of who issued the search, the mere fact that such a search was issued is extremely valuable information, as such, and the use of Tor is not enough to protect these kinds of queries.

While Tor does much to hide users' network information from the websites hosting content, it introduces a number of other problems. Nefarious Tor exit node operators have the ability to view, or worse, to modify the data that they relay. In at least one published case,[46] a server operator placed flash-based "webbugs" into webpages served in order to reveal the true source of the web requests. At the very least, an exit node operator has the ability to view users' search requests. Tor users must worry about the exit node operators keeping and later disclosing potentially sensitive searches in addition to search engine logging.

### B. INFORMATION LEAKAGE

It is commonly accepted among computer security experts that encrypting one's most sensitive communications is not enough, and in fact, can be extremely dangerous. The mere act of encrypting only sensitive messages leaks valuable information to outsiders watching the wire. They can see there are some messages that are important enough to try to protect. Techniques such as traffic analysis, when employed against a user who only encrypts important messages, can prove to be extremely effective.[47] Applying this idea to the problem of

---

[46] Andrew Christensen, "Practical Onion Hacking: Finding the Real Address of Tor Clients," *FortConsult*, 2006, http://www.packetstormsecurity.org/0610-advisories/Practical_Onion_Hacking.pdf (accessed November 8, 2007).

[47] George Danezis, "Introducing Traffic Analysis: Attacks, Defences and Public Policy Issues. . ." (lecture, 23rd Chaos Communication Congress (23C3), Berlin, Germany,

sensitive searches, it is quite clear that to achieve privacy one must apply the protection methods to all searches, and not just those that one deems to be sensitive.

Many privacy enhancing technologies impose rather steep costs on the user, such as lack of convenience due to the absence of cookie tracking across sessions or in the case of Tor, a significant increase in traffic latency as encrypted packets bounce across the globe before they reach their final destination. While users may be willing to tolerate this in order to gain privacy protections for sensitive searches, they may be less willing to do so for the bulk of their less sensitive traffic. This selective use of Tor and other technologies will unfortunately leave users vulnerable to traffic analysis by those with wiretap or network level access to users' communication data.

### C. TRAFFIC ANALYSIS AND PORNOGRAPHY

It has been noted by some observers that pornography drives technology.[48] Consumers of pornography are often the early adopters of technology and are often willing to tolerate beta quality products that other users refuse to use. One example of this is the initial attempts to stream video on the web. Users of adult content were the main audience willing to accept excruciatingly slow downloads of jittery, low quality videos. These early adopters wanted better video quality, and their demand arguably drove the market to develop better technology that eventually reached the masses.[49]

There is a notable absence of good information on the traffic that the Tor network carries, primarily because collecting such data in the United States could put researchers into legal jeopardy.[50] One recent study performed outside the United States suggests that one of the primary uses of Tor is to transfer pornographic content.[51] If the anecdotal evidence presented in this report accurately describes the

---

December 30, 2006) http://homes.esat.kuleuven.be/~gdanezis/TAIntro.pdf (accessed November 8, 2007).

[48] Peter Johnson, "Pornography Drives Technology: Why *Not* to Censor the Internet," FED. COMM. L. J., 49, no. 1 (1996): 217, http://www.law.indiana.edu/fclj/pubs/v49/no1/johnson.html (accessed November 8, 2007).

[49] Ibid.

[50] Electronic Frontier Foundation, "Tor: Legal FAQ for Tor Server Operators," April 25, 2005, http://tor.eff.org/eff/tor-legal-faq.html (accessed November 8, 2007).

[51] Christensen, "Practical Onion Hacking."

Tor network, it is thus likely that Tor traffic has a higher porn-per-packet ratio than "normal" Internet data. Given the assumption that a Tor user is probably interested in adult content, Google could allow advertisers to bid on keywords displayed to users coming from Tor exit nodes. "Tor targeting" would surely seem valuable to pornographic advertisers and would be a way for them to guess user intent without knowing anything else about a market segment that fiercely guards its privacy. This is just one example of the kind of user data that could be inferred from the use of a privacy-preserving system, even when encryption is used.

## VI. OTHER OPTIONS

TrackMeNot and Tor are not enough to protect vanity searches. This article now explores a few other options.

### A. SEARCHING ON ENCRYPTED DATA AND PRIVATE INFORMATION RETRIEVAL

There has been a significant amount of research into the areas of searching on encrypted data[52] and private information retrieval (PIR).[53] Search engines are focused on collecting more customer information, not in protecting consumers' privacy. Their business models are built on the practice of mining and exploiting user online behavior data. Thus, while research in these areas is interesting from an academic perspective, there is very little incentive for a service provider to expend the significant resources required to support PIR or

---

[52] Dawn Xiaodong Song, David Wagner, and Adrian Perrig, "Practical Techniques for Searches on Encrypted Data," in *Proceedings of the 2000 IEEE Symposium on Security and Privacy* (Washington, DC: IEEE Computer Society, 2000) http://www.cs.berkeley.edu/~dawnsong/papers/se.pdf (accessed November 8, 2007); Dan Boneh and others, "Public Key Encryption with Keyword Search," in *Advances in Cryptology – EUROCRYPT 2004* (Heidelberg, Germany: Springer-Berlin, 2004): 506–522, http://crypto.stanford.edu/~dabo/papers/encsearch.pdf (accessed November 8, 2007); Philippe Golle, Jessica Staddon, and Brent Waters, "Secure Conjunctive Keyword Search Over Encrypted Data," in *Applied Cryptography and Network Security* (Heidelberg, Germany: Springer-Berlin, 2004): 31–45, http://crypto.stanford.edu/~pgolle/papers/conj.pdf (accessed November 8, 2007).

[53] Amos Beimel and others, "Breaking the O(n1/(2k-1)) Barrier for Information-Theoretic Private Information Retrieval," in *Proceedings of the 43rd Symposium on Foundations of Computer Science* (Washington, DC: IEEE Computer Society, 2002): 261–70; Benny Chor and others, "Private Information Retrieval," in *Proceedings of the 36th Annual Symposium on Foundations of Computer Science* (Washington, DC: IEEE Computer Society, 1995): 41–50.

encrypted data searching, especially given that these technologies would hinder their primary goal of collecting customer data. The burden of privacy protection thus falls on the user.

### B. DOING THE SEARCH YOURSELF

Local searching, a surprisingly simple technique, may prove to be extremely useful at maintaining user privacy online. For fairly uncommon names, users can simply request every single web page containing their name or online nickname from a search engine, preferably, using an anonymizing proxy such as Tor. They can then download a copy of each of these web pages to their own computer, and perform a local search on those web pages for the sensitive terms.

This method has several shortcomings. A person with a unique name, but a fairly major web presence, may find that there are far too many web pages citing his or her name to download. Likewise, someone with a common name may encounter too many false positives when attempting to save a local copy of every page referencing him or her. In both of these cases, a complete local search may prove to be impossible. Finally, while most major operating systems include the ability to search through a large number of directories and files for one or more phrases, in many cases, it is not easy to use. The technology required to download every instance of a user's name from the Internet requires automation software, something not readily available to the masses. Thus, effective local searching is not yet an option for the vast majority of users.

Local searching lacks the bells, whistles, and ease of use that Google and the other search engines provide. Yet, it remains far safer in terms of user privacy than sending a sensitive vanity search out onto the Internet.

### C. PRE-ANNOUNCING YOUR STRATEGY

Technologies such as TrackMeNot pose a specific threat to the advertising dependant search engines. Instead of merely free-riding on Google's network and computing resources, as do those who search without viewing ads, TMN's method of achieving anonymity has the potential to cause significant collateral damage (via click-fraud) to Google's advertising system by requesting web pages with advertisements that will never be clicked. While TMN's goals are noble, its methods can cause unintentional harm to Google and others. One simple technique that could solve this problem of collateral damage would be for TMN users to disclose their intentions ahead of

time. By marking all search requests, both genuine and automated, with an additional argument in the query sent to Google's servers, TMN users could significantly reduce Google's incentive to locate and neutralize TMN traffic.[54] All queries originating from TMN users could then be easily excluded from the advertising system. While TMN's network activity probably will not be too difficult to differentiate from real user traffic, this simple technique at least reduces Google's incentive to do so. Just as webmasters can currently include a "robots.txt" file on their websites to notify web-crawlers of their desire to not be crawled, adding an additional flag to the search query would be a polite and reasonable way for TMN and other privacy preserving systems to communicate intent to Google.

The downside to this flagging technique is that by adding the flag, a user instantly announces himself as a TMN user. This then reduces his anonymity set to that of all TMN users, a small minority of all search clients.[55] Conversely, without the flag, the user could potentially be any of Google's millions of search users. TMN's current behavior is anything but covert, and thus, a user has probably already reduced his anonymity set by using TMN, even if he has not explicitly announced it.

### D. BE YOUR OWN PROXY

As others have noted, anonymity loves company.[56] Users gain privacy and plausible deniability when they can blend into a large crowd of other users. The use of TrackMeNot can be summed up as: "Google knows who I am, but if I send enough fake queries, the

---

[54] Something similar to this is already done for Google search queries issued via the built-in search bar in the Firefox web browser. Such queries from the browser are tagged, so that the search engine can later give Mozilla a cut of any advertising revenue generated by the query. The Mozilla Corporation, which owns and develops Firefox, made over $56 million dollars in 2006 through its profit sharing arrangement with Google. See Mitchell Baker, "Beyond Sustainability," *Mitchell's Blog*, October 22, 2007, http://weblogs.mozillazine.org/mitchell/archives/2007/10/beyond_sustainability.html (accessed November 8, 2007).

[55] Andreas Pfitzmann and Marit Hansen, "Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management—A Consolidated Proposal for Terminology," May 29, 2006, http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.28.pdf (accessed November 8, 2007).

[56] Roger Dingledine and Nick Mathewson, "Anonymity Loves Company: Usability and the Network Effect," in *Proceedings of the Fifth Workshop on the Economics of Information Security* (Cambridge, UK: WEIS, 2006) http://freehaven.net/doc/wupss04/usability.pdf (accessed November 8, 2007).

company won't know which searches are real, and which are not." Tor and other anonymizing proxies instead adopt the philosophy of "if I can keep my network location secret from Google, then while it will know exactly which searches are being issued, it won't know who is initiating them." Additionally, Tor users not only reveal their search information to Google, but also reveal it to the operator of a Tor exit node, who might not be trustworthy. One way of avoiding this problem of revealing search data to proxy operators is for users to run their own Tor exit node. Typically, when using the Tor network, users risk nefarious exit-node operators seeing their search queries. By assuming the role of an exit-node operator and using their own exit node for queries to Google and the other search engines, a user can make it far more difficult for another proxy administrator to learn of his or her search data.[57]

## VII. CONCLUSION

This article explored the problem of sensitive information leakage due to vanity searches on the Internet. It highlighted the inherent conflict of interest in the advertising/search engine business in which Google's ability to serve fine-grained advertising (and thus achieve higher revenues) directly competes with the methods by which users can achieve anonymity with the goal of preserving what little is left of their online privacy. This article also highlighted the state of information asymmetry between the search engines and users that makes it almost impossible to create artificial search queries that are indistinguishable from those submitted by real users. This article demonstrated that technologies such as TrackMeNot may increase user exposure through their attempts to create cover traffic than if they had not been used in the first place. Finally, this article explained how anonymizing proxies such as Tor are not enough to protect vanity searches. Several other potential solutions were discussed, none of which are ideal or completely foolproof.

Effective privacy protection for vanity searches is a difficult problem. Current privacy-preserving systems, although appearing to solve the problem, may only exacerbate it. Existing technologies alone cannot be trusted to provide private searching functionality to

---

[57] The Tor Project, "Do I Get Better Anonymity if I Run a Relay? Yes, You Do Get Better Anonymity Against Some Attacks," *Tor FAQ*, October 21, 2007, http://wiki.noreply.org/noreply/TheOnionRouter/TorFAQ (accessed November 8, 2007).

users. While still an unresolved issue, future research in the area will hopefully fill this dire need for privacy-preserving vanity searches.